A Tight Version of the Gaussian min-max theorem in the Presence of Convexity

Christos Thrampoulidis, Samet Oymak and Babak Hassibi
Department of Electrical Engineering, Caltech, Pasadena

Abstract

Gaussian comparison theorems are useful tools in probability theory; they are essential ingredients in the classical proofs of many results in empirical processes and extreme value theory. More recently, they have been used extensively in the analysis of underdetermined linear inverse problems. A prominent role in the study of those problems is played by *Gordon's Gaussian min-max theorem*. It has been observed that the use of the Gaussian min-max theorem produces results that are often *tight*. Motivated by recent work due to M. Stojnic, we argue explicitly that the theorem is tight under additional *convexity* assumptions. To illustrate the usefulness of the result we provide an application example from the field of noisy linear inverse problems.

I. Introduction

Proposition I.1 below is an important variation of the *Gaussian min-max theorem* proved by Gordon in [1]. The version that we present here is only a slightly modified version of the original result as it appears in [1, Lemma 3.1][1]. For completeness, we include some background and a proof of Proposition I.1 in Appendix A.

Proposition I.1 (Gaussian min-max theorem (GMT))**.** *Let $\mathbf{A}\in\mathbb{R}^{m\times n}$, $g\in\mathbb{R}$, $\mathbf{g}\in\mathbb{R}^m$ and $\mathbf{h}\in\mathbb{R}^n$ have entries i.i.d. $\mathcal{N}(0,1)$ and be independent of each other. Also, let $\mathcal{S}_1\subset\mathbb{R}^n$, $\mathcal{S}_2\subset\mathbb{R}^m$ be compact sets and $\psi(\cdot,\cdot)$ be a continuous function on $\mathcal{S}_1\times\mathcal{S}_2$. Finally, define*[2]

$$\mathcal{F}(\mathbf{A},g):=\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\ \mathbf{y}^T\mathbf{A}\mathbf{x}+g\|\mathbf{x}\|_2\|\mathbf{y}\|_2+\psi(\mathbf{x},\mathbf{y}), \tag{1}$$

$$\mathcal{G}(\mathbf{g},\mathbf{h}):=\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\ \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y}+\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x}+\psi(\mathbf{x},\mathbf{y}). \tag{2}$$

Then, for any $c\in\mathbb{R}$:

$$\mathbb{P}\left(\mathcal{F}(\mathbf{A},g)<c\right)\leq\mathbb{P}\left(\mathcal{G}(\mathbf{g},\mathbf{h})\leq c\right). \tag{3}$$

Consider c such that $\mathbb{P}\left(\mathcal{G}(\mathbf{g},\mathbf{h})\leq c\right)$ is close to zero; we say that c is a high-probability *lower* bound to $\mathcal{G}(\mathbf{g},\mathbf{h})$. According to Proposition I.1, c is also a high-probability lower bound to $\mathcal{F}(\mathbf{A},g)$. In that sense, Proposition I.1 can be used as a tool to derive high-probability lower bounds on $\mathcal{F}(\mathbf{A},g)$; this is achieved indirectly via analyzing the different optimization problem defined in (2), which we will frequently refer to as "*Gordon's optimization*". In many interesting cases, the analysis of the latter is much easier to perform[3]. We refer the reader to [1] and Section IV for specific applications of this idea.

A natural question that arises concerns the *tightness* of the bounds obtained from Proposition I.1. To be more explicit, suppose that the following concentration inequality holds[4] for $\mathcal{G}(\mathbf{g},\mathbf{h})$. There exists $L>0$ such that for all $t>0$, the events

$$\{\mathcal{G}(\mathbf{g},\mathbf{h})\leq\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})-t\}\qquad\text{and}\qquad\{\mathcal{G}(\mathbf{g},\mathbf{h})\geq\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})+t\},$$

each occurs with probability no larger than $\exp(-t^2/(2L^2))$. Then, $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})-t$ is a high-probability lower bound to $\mathcal{G}(\mathbf{g},\mathbf{h})$. This bound is also *tight* in the sense that it is accompanied by a corresponding high-probability upper bound, namely $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})+t$, whose value can be made arbitrarily close to the former. Now, Proposition I.1 tells us that $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})-t$ is also a high-probability lower bound on $\mathcal{F}(\mathbf{A},g)$. Yet, it gives *no* information on how *tight* this bound is.

In this note, we show that under additional *convexity* assumptions on the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the function $\psi(\cdot,\cdot)$, Proposition I.1 is tight in the sense discussed above. We essentially prove that in the presence of convexity GMT can be used to prove a counterpart for itself (i.e. to (3)) which *upper* bounds $\mathcal{F}(\mathbf{A},g)$ in terms of $\mathcal{G}(\mathbf{g},\mathbf{h})$. Our result is motivated by recent work of Stojnic [2, 3, 4]. We discuss connection to this line of work in Section V.

The rest of the paper is organized as follows. In Section II we state our main result Theorem II.1 and include its proof in Section III. We illustrate the usefulness of our result through a specific application in Section IV. We conclude the paper in Section V with a discussion on relevant work.

[1] In contrast to Proposition I.1, Lemma 3.1 in [1] assumes $\mathcal{S}_1$ to be arbitrary (not necessarily compact) subset of $\mathbb{R}^m$, $\mathcal{S}_2$ is restricted to be the unit sphere in $\mathbb{R}^n$ and $\psi(\cdot,\cdot)$ is only a function of $\mathbf{x}$.

[2] Although not explicit in the definition, it should be clear that $\mathcal{F}(\mathbf{A},g)$ and $\mathcal{G}(\mathbf{g},\mathbf{h})$ also depend on the particular choices of the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and of the function $\psi(\cdot,\cdot)$.

[3] Moving from (1) to (2) the term $\mathbf{y}^T\mathbf{A}\mathbf{x}$ is "decoupled" into two separate terms that involve a random gaussian vector each and are also independent of each other. The number of random variables is reduced from $mn+1$ to $m+n$.

[4] Lemma B.2 shows that $\mathcal{G}(\mathbf{g},\mathbf{h})$ is an L-Lipschitz function of $(\mathbf{g},\mathbf{h})$. The concentration result then follows from the Gaussian concentration of measure phenomenon for Lipschitz functions (Proposition B.1).

II. Main Result

Our main result is stated in Theorem II.1 below.

Theorem II.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$ have entries i.i.d. $\mathcal{N}(0,1)$ and be independent of each other. Also, let $\mathcal{S}_1 \subset \mathbb{R}^n$, $\mathcal{S}_2 \subset \mathbb{R}^m$ be nonempty compact sets and $\psi(\cdot,\cdot)$ be a continuous function on $\mathcal{S}_1 \times \mathcal{S}_2$. Finally, define*

$$\mathcal{F}(\mathbf{A}) := \min_{\mathbf{x}\in\mathcal{S}_1} \max_{\mathbf{y}\in\mathcal{S}_2} \ \mathbf{y}^T\mathbf{A}\mathbf{x} + \psi(\mathbf{x},\mathbf{y}), \tag{4}$$

$$\mathcal{G}(\mathbf{g},\mathbf{h}) := \min_{\mathbf{x}\in\mathcal{S}_1} \max_{\mathbf{y}\in\mathcal{S}_2} \ \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y}). \tag{5}$$

Then, for any $c_- \in \mathbb{R}$:

$$\mathbb{P}\left(\mathcal{F}(\mathbf{A}) < c_-\right) \leq 2\mathbb{P}\left(\mathcal{G}(\mathbf{g},\mathbf{h}) \leq c_-\right). \tag{6}$$

If in addition both the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are convex and $\psi(\cdot,\cdot)$ is convex-concave on $\mathcal{S}_1 \times \mathcal{S}_2$, then, for any $c_+ \in \mathbb{R}$:

$$\mathbb{P}\left(\mathcal{F}(\mathbf{A}) > c_+\right) \leq 2\mathbb{P}\left(\mathcal{G}(\mathbf{g},\mathbf{h}) \geq c_+\right). \tag{7}$$

Let us compare Theorem II.1 to Gordon's original result Proposition I.1. First, notice the slight difference in the optimization problems involved in the definitions (1) and (4); in contrast to Proposition I.1, the minimax optimization in (4) does not include the term "$g\|\mathbf{x}\|_2\|\mathbf{y}\|_2$". The "price" paid for this, is the multiplicative factor of 2 in (6), when compared to (3). Note however that this factor does not affect the essence of the result since the scenarios of interest are those for which $\mathbb{P}(\mathcal{G}(\mathbf{g},\mathbf{h}) \leq c)$ is close to zero. What is more, in all the applications[5], where GMT is useful, the optimization problem involved is in the form of (4) rather than that of (1). One reason behind this, is that under convexity assumptions on $\mathcal{S}_1$, $\mathcal{S}_2$ and $\psi(\cdot,\cdot)$ the minimax optimization in (4) is a *convex* program, which is generally more likely to be encountered in applications compared to the always non-convex program in (1). Convexity, is also critical for establishing the second statement of the theorem, namely inequality (7).

Inequality (6) is essentially no different than what Proposition I.1 states; if c_- is a high probability lower bound for Gordon's optimization $\mathcal{G}(\mathbf{g},\mathbf{h})$, so it is for $\mathcal{F}(\mathbf{A})$. The main contribution of Theorem II.1 is inequality (7). This holds only under imposing appropriate convexity assumption and provides a counterpart to (6) and GMT; if c_+ is a high probability *upper* bound for Gordon's optimization $\mathcal{G}(\mathbf{g},\mathbf{h})$, so it is for $\mathcal{F}(\mathbf{A})$.

Making a connection to our discussion in the introduction, Theorem II.1 shows that in the presence of convexity, GMT is tight. To see this apply the theorem for $c_+ = \mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h}) + t$ and $c_- = \mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h}) - t$. It is shown in Lemma B.2 in the Appendix that $\mathcal{G}(\mathbf{g},\mathbf{h})$ is Lipschitz in $(\mathbf{g},\mathbf{h})$. It then follows from Proposition B.1 and the Gaussian concentration of Lipschitz functions that $\mathcal{G}(\mathbf{g},\mathbf{h})$ shows normal concentration around its mean $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})$. Thus, we obtain Corollary II.1 below which shows that $\mathcal{F}(\mathbf{A})$ concentrates around $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})$.

Corollary II.1. *Let all the assumptions of Theorem II.1 hold. Further, define $R_1 := \max_{\mathbf{x}\in\mathcal{S}_1}\|\mathbf{x}\|_2$ and $R_2 := \max_{\mathbf{y}\in\mathcal{S}_2}\|\mathbf{y}\|_2$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\ |\mathcal{F}(\mathbf{A}) - \mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h})| > t\ \right) \leq 4\exp\left(-t^2/(4R_1^2R_2^2)\right).$$

As discussed in the introduction, the premise of GMT is that Gordon's optimization is significantly easier to analyze than the original quantity of interest $\mathcal{F}(\mathbf{A})$. In that sense, Theorem II.1 and in particular Corollary II.1 provide a powerful tool for proving tight probabilistic lower and upper bounds on $\mathcal{F}(\mathbf{A})$. In Section IV we illustrate how Corollary II.1 can be used to pinpoint the optimal cost of the LASSO algorithm.

III. Proof Of Theorem II.1

As discussed inequality (6) is an (almost) direct consequence of Gordon's Proposition I.1; to prove (7) we apply strong duality and appropriately apply GMT to the dual problem. But, the first critical step is to get rid of the term "$g\|\mathbf{x}\|_2\|\mathbf{y}\|_2$" in (3) in Gordon's Lemma I.1; this is summarized in Lemma III.1, below[6].

Lemma III.1. *Let the same assumptions as in the statement of Proposition I.1 hold. Then,*

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\ \left\{\mathbf{y}^T\mathbf{A}\mathbf{x} + \psi(\mathbf{x},\mathbf{y})\right\} < c\right) \leq 2\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y})\right\} \leq c\right). \tag{8}$$

Proof: Fix $\mathbf{A}$ and $g < 0$ and denote

$$f_1(\mathbf{x},\mathbf{y}) = \mathbf{y}^T\mathbf{A}\mathbf{x} + \psi(\mathbf{x},\mathbf{y}) \quad \text{and} \quad f_2(\mathbf{x},\mathbf{y}) = \mathbf{y}^T\mathbf{A}\mathbf{x} + g\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \psi(\mathbf{x},\mathbf{y}).$$

[5] e.g. [1, 4, 5] and Section IV and references therein.

[6] Lemma III.1 and Proposition I.1 first appear in the authors' earlier work [6].

Clearly, $f_1(\mathbf{x},\mathbf{y}) \geq f_2(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \mathcal{S}_1 \times \mathcal{S}_2$. We may then write,

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2} f_1(\mathbf{x},\mathbf{y}) = f_1(\mathbf{x}_1,\mathbf{y}_1) &\geq f_1(\mathbf{x}_1,\mathbf{y}) \text{ for all } \mathbf{y}\in\mathcal{S}_2\\ &\geq \max_{\mathbf{y}\in\mathcal{S}_2} f_2(\mathbf{x}_1,\mathbf{y}) \geq \min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2} f_2(\mathbf{x},\mathbf{y}).\end{aligned}$$

Recalling the definitions of $\mathcal{F}(\mathbf{A},g)$ and $\mathcal{F}(\mathbf{A})$ in (1) and (4), we have shown that $\mathcal{F}(\mathbf{A}) \geq \mathcal{F}(\mathbf{A},g)$. From this and from independence of g and $\mathbf{A}$, for all $c \geq 0$:

$$\mathbb{P}\left(\mathcal{F}(\mathbf{A},g) \geq c \mid g<0\right) \leq \mathbb{P}\left(\mathcal{F}(\mathbf{A}) \geq c \mid g<0\right) = \mathbb{P}(\mathcal{F}(\mathbf{A}) \geq c).$$

Using this and $g \sim \mathcal{N}(0,1)$, we find

$$\mathbb{P}(\mathcal{F}(\mathbf{A},g) \geq c) = \frac{1}{2}\mathbb{P}(\mathcal{F}(\mathbf{A},g) \geq c \mid g>0) + \frac{1}{2}\mathbb{P}(\mathcal{F}(\mathbf{A},g) \geq c \mid g<0) \leq \frac{1}{2} + \frac{1}{2}\mathbb{P}(\mathcal{F}(\mathbf{A}) \geq c).$$

We may now apply Gordon's Proposition I.1 to conclude with

$$\begin{aligned}\mathbb{P}(\mathcal{F}(\mathbf{A}) \geq c) &\geq 2\mathbb{P}(\mathcal{F}(\mathbf{A},g) \geq c) - 1\\ &\geq 2\mathbb{P}(\mathcal{G}(\mathbf{h},\mathbf{g}) > c) - 1. \end{aligned} \tag{9}$$

It is now straightforward to conclude with (8). ■

To prove (6) just apply Lemma III.1. Observe that this did not require any convexity assumptions.

In what follows we prove (7). By assumption, the sets $\mathcal{S}_1, \mathcal{S}_2$ are non-empty compact and convex. Furthermore, the function $\mathbf{y}^T\mathbf{A}\mathbf{x} + \psi(\mathbf{x},\mathbf{y})$ is continuous, finite[7] and convex-concave on $\mathcal{S}_1 \times \mathcal{S}_2$. Thus, we can apply the minimax result in [7, Corollary 37.3.2] to exchange "min-max" with a "max-min" in (5):

$$\mathcal{F}(\mathbf{A}) = \max_{\mathbf{y}\in\mathcal{S}_2}\min_{\mathbf{x}\in\mathcal{S}_1} \mathbf{y}^T\mathbf{A}\mathbf{x} + \psi(\mathbf{x},\mathbf{y})$$

It is convenient to rewrite the above as

$$-\mathcal{F}(\mathbf{A}) = \min_{\mathbf{y}\in\mathcal{S}_2}\max_{\mathbf{x}\in\mathcal{S}_1} -\mathbf{y}^T\mathbf{A}\mathbf{x} - \psi(\mathbf{x},\mathbf{y}).$$

Then, using the symmetry of $\mathbf{A}$, we have that for any $c>0$:

$$\mathbb{P}\left(-\mathcal{F}(\mathbf{A}) \leq c\right) = \mathbb{P}\left(\min_{\mathbf{y}\in\mathcal{S}_2}\max_{\mathbf{x}\in\mathcal{S}_1}\left\{\mathbf{y}^T\mathbf{A}\mathbf{x} - \psi(\mathbf{x},\mathbf{y})\right\} \leq c\right)$$

We may now apply apply[8] Lemma III.1:

$$\begin{aligned}\mathbb{P}\left(-\mathcal{F}(\mathbf{A}) < c\right) &\leq 2\mathbb{P}\left(\min_{\mathbf{y}\in\mathcal{S}_2}\max_{\mathbf{x}\in\mathcal{S}_1}\left\{\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} - \psi(\mathbf{x},\mathbf{y})\right\} \leq c\right)\\ &= 2\mathbb{P}\left(\min_{\mathbf{y}\in\mathcal{S}_2}\max_{\mathbf{x}\in\mathcal{S}_1}\left\{-\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} - \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} - \psi(\mathbf{x},\mathbf{y})\right\} \leq c\right),\end{aligned} \tag{10}$$

where the equality follows because of the symmetry of $\mathbf{g}$ and $\mathbf{h}$. To continue, note that

$$\min_{\mathbf{y}\in\mathcal{S}_2}\max_{\mathbf{x}\in\mathcal{S}_1}\left\{-\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} - \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} - \psi(\mathbf{x},\mathbf{y})\right\} = -\max_{\mathbf{y}\in\mathcal{S}_2}\min_{\mathbf{x}\in\mathcal{S}_1}\left\{\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \psi(\mathbf{x},\mathbf{y})\right\},$$

and further apply the minimax inequality [7, Lemma 36.1] which requires that for all $\mathbf{g},\mathbf{h}$

$$\max_{\mathbf{y}\in\mathcal{S}_2}\min_{\mathbf{x}\in\mathcal{S}_1}\left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y})\right\} \leq \min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y})\right\} := \mathcal{G}(\mathbf{g},\mathbf{h}).$$

These, when combined with (10), give

$$\mathbb{P}\left(-\mathcal{F}(\mathbf{A}) < c\right) \leq 2\mathbb{P}\left(-\mathcal{G}(\mathbf{g},\mathbf{h}) \leq c\right).$$

Apply the above for $c = -c_+$, to conclude with (7), as desired.

[7] A continuous function on a compact set is bounded from Weierstrass extreme value theorem.

[8] Observe that in (8) the signs of the terms $\mathbf{y}^T\mathbf{A}\mathbf{x}$, $\mathbf{g}^T\mathbf{y}$ and $\mathbf{h}^T\mathbf{x}$ do not matter because of the assumed symmetry in the distributions of $\mathbf{A},\mathbf{g}$ and $\mathbf{h}$.

IV. Application

A. Motivation

We illustrate the usefulness of Theorem II.1 and Corollary II.1 through an example. We consider the task of estimating an unknown but *structured* signal $\mathbf{x}_0 \in \mathbb{R}^n$ from noisy linear observations $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$, where $\mathbf{A}$ is a measurement matrix and $\mathbf{z}$ is the noise vector. $\mathbf{x}_0$ is structured in the sense that it actually lives in a manifold of lower dimension than the dimension n of the ambient space . Typical examples of such signals include sparse, low-rank, block-sparse and many more (e.g. [8]). To promote the particular structure of $\mathbf{x}_0$ associate with it some appropriate structure-inducing *convex* function $f(\cdot)$. For example, if $\mathbf{x}_0$ is sparse then $f(\cdot)$ can be the ℓ_1-norm (see [8] for more examples). A reasonable estimate for $\mathbf{x}_0$ is obtained as the solution $\hat{\mathbf{x}}$ of the following program:

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\| \quad \text{s.t.} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0), \tag{11}$$

where $\|\cdot\|$ is some appropriately chosen norm. For example, if the ℓ_2-norm is chosen in the objective then the algorithm in (11) becomes the celebrated LASSO method in the statistics literature [9]. We wish to characterize the estimation performance of (11). In particular, we are interested in establishing tight upper bounds on the normalized squared error $\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2/\|\mathbf{z}\|_2^2$. Before proceeding it is convenient to rewrite (11) after changing the decision variable to $\mathbf{w} := \mathbf{x} - \mathbf{x}_0$. Also, we are interested in a "first-order analysis" of the problem and, thus, we will relax the constraint set $\mathcal{D}_f(\mathbf{x}_0) = \{\mathbf{v} \mid f(\mathbf{x}_0 + \mathbf{v}) \leq f(\mathbf{x}_0)\}$ to (essentially) its conic hull $\mathcal{T}_f(\mathbf{x}_0) = \mathrm{Cl}(\mathrm{cone}(\mathcal{D}_f(\mathbf{x}_0))$, where $\mathrm{Cl}(\cdot)$ denotes the set closure operator and $\mathrm{cone}(\cdot)$ returns the conic hull of a set. See [6] for details on this first-order approximation. With these the program in (11) becomes:

$$\hat{\mathbf{w}} = \min_{\mathbf{w} \in T_f(\mathbf{x}_0)} \|\mathbf{z} - \mathbf{A}\mathbf{w}\|. \tag{12}$$

Onwards we assume that the entries of $\mathbf{A}$ are i.i.d. standard normal. Further, assume that the optimal solution of (12) is bounded by some large constant $\|\hat{\mathbf{w}}\|_2 \leq C$; the value of C is to be specified later. Under these assumptions, our goal is to establish a sharp bound on the estimation error $\|\hat{\mathbf{w}}\|_2$.

The framework that we will use was first introduced by Stojnic in [10]. It turns out that a critical step in this proof technique is to pinpoint the optimal cost of the optimization in (12), call it $\mathcal{F}(\mathbf{A}, \mathbf{z})$[9]. This was accomplished by Stojnic in [10] for the LASSO algorithm when $f(\cdot) = \|\cdot\|_1$ and $\mathbf{x}_0$ is sparse and by the authors in [6] for arbitrary convex regularizers $f(\cdot)$. The idea behind this is to combine strong duality with Gordon's Lemma and is attributed to Stojnic. Of course, this is also the core of the idea behind Theorem II.1. However, the treatment in [10] and [6] is significantly involved and the proof of the result on $\mathcal{F}(\mathbf{A}, \mathbf{z})$ requires several pages. Here, with an appropriate slight modification to the technique introduced in [10] and applying Corollary II.1 we are able to reproduce the same result in a more principled and concise way. This approach also facilitates generalizations of the results of [10],[6] to norms other than the ℓ_2 in the objective of (12) (see [11]).

B. Theorem II.1 in use

We will assume that the noise vector $\mathbf{z} = \sigma\mathbf{v}$ where the entries of $\mathbf{v}$ are i.i.d. standard normal and σ^2 is the noise variance. We may equivalently express the optimal cost $\mathcal{F}(\mathbf{A}, \mathbf{z})$ of (12) as

$$\mathcal{F}(\mathbf{A}, \mathbf{g}) = \min_{\substack{\mathbf{w} \in T_f(\mathbf{x}_0) \\ \|\mathbf{w}\|_2 \leq C}} \max_{\|\mathbf{a}\|_2 \leq 1} \mathbf{a}^T \begin{bmatrix} A & -\mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \tag{13}$$

The minimax problem above satisfies all assumptions of Corollary II.1. In what follows we write Gordon's optimization problem corresponding to (13). First, let $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$, $h \in \mathbb{R}$ have i.i.d. standard normal entries. Also, denote $\mathcal{S}^{n-1}$ the unit sphere in $\mathbb{R}^n$ and $(\chi)_+ := \max\{\chi, 0\}$ for any $\chi \in \mathbb{R}$.

$$\begin{aligned}
\mathcal{G}(\mathbf{g}, \mathbf{h}, h) &= \min_{\substack{\mathbf{w} \in T_f(\mathbf{x}_0) \\ \|\mathbf{w}\| \leq C}} \max_{\|\mathbf{a}\|_2 \leq 1} \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\mathbf{a}^T\mathbf{g} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{w} - \|\mathbf{a}\|_2 h\sigma \\
&= \min_{\substack{\mathbf{w} \in T_f(\mathbf{x}_0) \\ \|\mathbf{w}\| \leq C}} \max_{0 \leq \beta \leq 1} \beta \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} - h\sigma \right) \\
&= \min_{\substack{\mathbf{w} \in T_f(\mathbf{x}_0) \\ \|\mathbf{w}\| \leq C}} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} - h\sigma \right)_+ \\
&= \left(\min_{0 \leq \alpha \leq C} \left\{ \sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \alpha \max_{\mathbf{w} \in \mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{S}^{n-1}} \mathbf{h}^T\mathbf{w} \right\} + h\sigma \right)_+ .
\end{aligned} \tag{14}$$

[9] In short, Stojnic's idea [10] in proving a tight upper bound, say U, on $\|\hat{\mathbf{w}}\|_2$ is to prove that for any $\epsilon > 0$: $\mathbb{P}(\min_{\substack{\mathbf{w} \in T_f(\mathbf{x}_0) \\ (1+\epsilon)U \leq \|\mathbf{w}\| \leq C}} \|\mathbf{z} - \mathbf{A}\mathbf{w}\| > \mathcal{F}(\mathbf{A}, \mathbf{z})) \to 1$.

For convenience, define the function $d:\mathbb{R}^n\to\mathbb{R}$:

$$d(\mathbf{h}):=\max_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1}}\mathbf{h}^T\mathbf{w}.$$

The minimization problem in (14) is a rather simple scalar optimization. It takes not much effort to show that[10] (see Lemma C.2) when $\|\mathbf{g}\|_2>d(\mathbf{h})>0$, then its optimal cost is $\sigma\sqrt{\|\mathbf{g}\|_2^2-d(\mathbf{h})^2}$. Furthermore, both functions $\|\cdot\|_2$ and $d(\cdot)$ are 1-Lipschitz (see for example [1]). Thus, using the Gaussian concentration of measure (Proposition (B.1)) they nicely concentrate around their means, which we denote:

$$\gamma_m:=\mathbb{E}\|\mathbf{g}\|_2\quad\text{ and }\quad\omega:=\omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1}):=\mathbb{E}d(\mathbf{h}).$$

It is well known that $\frac{m}{\sqrt{m+1}}\le\gamma_m\le\sqrt{m}$ and $\omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1})$ is known as the "gaussian width" [11] of $\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{S}^{n-1}$. In Appendix C we use those and similar ideas to yield an expression for $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)$. In particular, we prove in Lemma C.1 that when $\gamma_m>\omega$, then for any $0<\delta<1$, there exists sufficiently large m such that

$$(1-\delta)\sigma\sqrt{\gamma_m^2-\omega^2}\le\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)\le(1+\delta)\sigma\sqrt{\gamma_m^2-\omega^2}.\tag{15}$$

Now we are ready to apply Corollary II.1 to translate this into a tight concentration result for the objective of the LASSO $\mathcal{F}(\mathbf{A},\mathbf{z})$.

Theorem IV.1. *Assume $0<\epsilon<1$ such that $(1-\epsilon)\gamma_m>\omega>\epsilon\gamma_m$ and m is large enough. Recall the definition of $\mathcal{F}(\mathbf{A},\mathbf{z})$ in (13) and, therein, let $C>\sigma\omega/\sqrt{\gamma_m^2-\omega^2}$. Define $G_*:=\sigma\sqrt{\gamma_m^2-\omega^2}$. Then, for all $0<\delta<1$, there exists constant $c:=c(\delta,\sigma,C,\epsilon)$ such that with probability $1-e^{-c\gamma_m^2}$:*

$$(1-\delta)G_*\le\mathcal{F}(\mathbf{A},\mathbf{z})\le(1+\delta)G_*.$$

Proof: Fix any $0<\delta<1$ and let $\eta=\sqrt{1+\delta}-1$. First, apply Corollary II.1;in our case $R_1=C$ and $R_2=1$. Hence,

$$\mathbb{P}\left(\ |\mathcal{F}(\mathbf{A},\mathbf{z})-\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)|>\eta\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)\ \right)\le4\exp\left(-\eta^2(\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h))^2/(4C^2)\right).\tag{16}$$

Next, let m be sufficiently large such that (15) holds for η. Combining this with (16) we conclude with the desired,

$$\mathbb{P}\left(\ |\mathcal{F}(\mathbf{A},\mathbf{z})-G_*|>\delta G_*\ \right)\le4\exp\left(-\sigma^2(1-\eta)^2\eta^2\epsilon(2-\epsilon)\gamma_m^2/(4C^2)\right).$$

■

V. Related Work and Conclusion

Starting from the work of Vershynin and Rudelson [16], Gaussian comparison theorems have played instrumental role in developing a clear understanding of linear inverse problems when the measurement matrix follows the standard Gaussian distribution. The idea of combining strong duality with the Gaussian min-max theorem (GMT) is originally attributed to Stojnic [2]. In a recent line of work he makes repeated use of this powerful idea. In [3] he applies it to prove that the ℓ_1-minimization phase transition thresholds of [8, 12] are tight. A similar observation also appears in [13] by Amelunxen et.al.. In these works, the strong duality argument originates from the KKT optimality conditions rather than swapping min-max. Furthermore, Stojnic applies this idea to prove a tight upper bound on the normalized squared error of the LASSO algorithm with ℓ_1 regularization [10]. The result was later generalized and extended in various directions by the current authors in [6, 17, 18]. Finally, Stojnic showed how similar ideas can be applied to the study of the storage capacity of perceptrons [4].

This work is motivated by and builds upon Stojnic's idea. Our insights and additional technical effort lead to a succinct statement of our main result in Theorem II.1 and Corollary II.1, which both appear to be novel. In Theorem II.1 we have quantified the exact (convexity) conditions that are required for a counterpart to the inequality in Proposition I.1 to hold true. A critical observation amounts to the fact that through a symmetrization trick we can get rid of the term $g\|\mathbf{x}\|_2\|\mathbf{y}\|_2$ in one of the Gaussian processes involved in GMT. The resulting minimax optimization problem is now convex and the rest follows. The message of Corollary II.1 is simple: the two minimax optimization problems introduced in Proposition I.1 are such that the first concentrates around the mean of the second. In that sense, we have shown explicitly that when combined with convexity GMT is tight. In Section IV we showed the power of Corollary II.1 by applying it to pinpoint the optimal cost of the LASSO optimization. In particular, we were able to recover a result from [6] with substantially less effort and through a more insightful treatment. The direct and simplified nature of Corollary II.1, when compared to the rather complex arguments in [6, 10], allows for extensions to other problems than the LASSO. Our work in [11], that analyzes the constrained least absolute deviations algorithm (problem (11) with $\|\cdot\|=\|\cdot\|_1$), is an example towards this direction.

[10]here, assume C large enough. See Appendix C for the exact statements.

[11] The gaussian width ω appears as a fundamental quantity in the study of noiseless Compressed Sensing, where one wishes to recover an unknown structured signal $\mathbf{x}_0\in\mathbb{R}^n$ from $m<n$ linear equations via $\min f(\mathbf{x})$ s.t. $\mathbf{A}\mathbf{x}=\mathbf{A}\mathbf{x}_0$. Earlier works [8, 12] had proved that $m>\omega^2$ number of measurements suffice for this convex algorithm to uniquely recover $\mathbf{x}_0$. More recently, it was shown independently in [3, 13] that ω^2 number of measurements are also necessary for unique recovery. The arguments in [3] rely on GMT, while [13] uses tools from conic integral geometry; see [14] for a connection between those two. It is important to note that the gaussian width ω admits accurate estimates for a number of important regularizers $f(\cdot)$. For example, for $f(\cdot)=\|\cdot\|_1$ and $\mathbf{x}_0$ k-sparse, it is shown in [8, 13] that $\omega^2\lesssim2k\log(n/k)+(3/2)k$. See [8, 13, 15] for more examples.

REFERENCES

[1] Y. Gordon, *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.
[2] M. Stojnic, "Meshes that trap random subspaces," *arXiv preprint arXiv:1304.0003*, 2013.
[3] ——, "Upper-bounding ℓ_1-optimization weak thresholds," *arXiv preprint arXiv:1303.7289*, 2013.
[4] ——, "Spherical perceptron as a storage memory with limited errors," *arXiv preprint arXiv:1306.3809*, 2013.
[5] ——, "Bounding ground state energy of hopfield models," *arXiv preprint arXiv:1306.3764*, 2013.
[6] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized lasso: A precise analysis," *arXiv preprint arXiv:1311.0830*, 2013.
[7] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1997, vol. 28.
[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.
[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
[10] M. Stojnic, "A framework to characterize performance of lasso algorithms," *arXiv preprint arXiv:1303.7291*, 2013.
[11] C. Thrampoulidis and B. Hassibi, "Estimating structured signals in sparse noise: A precise noise sensitivity analysis," in *52nd Annual Allerton Conference (to appear)*. IEEE, 2014.
[12] M. Stojnic, "Various thresholds for ℓ_1-optimization in compressed sensing," *arXiv preprint arXiv:0907.3666*, 2009.
[13] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: A geometric theory of phase transitions in convex optimization," *arXiv preprint arXiv:1303.6672*, 2013.
[14] D. Amelunxen and M. Lotz, "Gordon's inequality and condition numbers in conic optimization," *arXiv preprint arXiv:1408.3016*, 2014.
[15] R. Foygel and L. Mackey, "Corrupted sensing: Novel guarantees for separating structured signals," *arXiv preprint arXiv:1305.2524*, 2013.
[16] M. Rudelson and R. Vershynin, "Sparse reconstruction by convex relaxation: Fourier and gaussian measurements," in *Information Sciences and Systems, 2006 40th Annual Conference on*. IEEE, 2006, pp. 207–212.
[17] S. Oymak, C. Thrampoulidis, and B. Hassibi, "Simple bounds for noisy linear inverse problems with exact side information," *arXiv preprint arXiv:1312.0641*, 2013.
[18] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Simple error bounds for regularized noisy linear inverse problems," *Information Theory, 2014. ISIT 2014. Proceedings. International Symposium on*, pp. 3007–3011, 2014.
[19] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991, vol. 23.
[20] Y. Gordon, "Some inequalities for gaussian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.
[21] ——, "Elliptically contoured distributions," *Probability theory and related fields*, vol. 76, no. 4, pp. 429–438, 1987.
[22] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
[23] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

APPENDIX

A. Gordon's Gaussian min-max theorem

Gaussian comparison theorems are powerful tools in probability theory [19]. A particularly useful such comparison inequality is described by Gordon's comparison theorem. In fact Gordon's theorem, is a generalization of the classical Slepian lemma and Fernique theorem [20]. It was first proved by Y. Gordon in [20], where it was also shown how it can be used as an alternative to (re)-derive other well-known results in the field. See also [21] for slight generalized versions of the theorem and the classical reference [19, Chapter 3.3] for an introduction to gaussian comparison theorems some applications.

Theorem A.1 (Gordon's Gaussian comparison theorem, [20])**.** *Let* $\{X_{ij}\}$ *and* $\{Y_{ij}\}$, $1 \leq i \leq I$, $1 \leq j \leq J$, *be centered Gaussian processes such that*

$$\begin{cases} \mathbb{E}X_{ij}^2 = \mathbb{E}Y_{ij}^2, & \text{for all } i,j, \\ \mathbb{E}X_{ij}X_{ik} \geq \mathbb{E}Y_{ij}Y_{ik}, & \text{for all } i,j,k, \\ \mathbb{E}X_{ij}X_{\ell k} \leq \mathbb{E}Y_{ij}Y_{\ell k}, & \text{for all } i \neq \ell \text{ and } j,k. \end{cases}$$

Then, for all $\lambda_{ij} \in \mathbf{R}$,

$$\mathbb{P}\left(\bigcap_{i=1}^{I}\bigcup_{j=1}^{J}[Y_{ij} \geq \lambda_{ij}]\right) \geq \mathbb{P}\left(\bigcap_{i=1}^{I}\bigcup_{j=1}^{J}[X_{ij} \geq \lambda_{ij}]\right).$$

Gordon's Theorem A.1 establishes a probabilistic comparison between two abstract Gaussian processes $\{X_{ij}\}$ and $\{Y_{ij}\}$ based on conditions on their corresponding covariance structures. Proposition I.1 is a corollary of Theorem A.1 when applied to specific Gaussian processes.

We begin with using Theorem A.1 to prove an analogue of Proposition I.1 for discrete sets. The proof is almost identical to the proof of Gordon's original Lemma 3.1 in [1]. Nevertheless, we include it here for completeness. After the proof of Lemma A.1, we use a compactness argument to translate the result to continuous sets and complete the proof of Proposition I.1.

To simplify notation we suppress notation and write $\|\cdot\|$ instead of $\|\cdot\|_2$.

Lemma A.1 (Gordon's Gaussian min-max theorem: Discrete Sets)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $g \in \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$ have entries i.i.d. $\mathcal{N}(0,1)$ and be independent of each other. Also, let $\mathcal{I}_1 \subset \mathbb{R}^n$, $\mathcal{I}_2 \subset \mathbb{R}^m$ be finite sets of vectors and $\psi(\cdot,\cdot)$ be a finite function defined on $\mathcal{I}_1 \times \mathcal{I}_2$. Then, for all $c > 0$:*

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{I}_1}\max_{\mathbf{y}\in\mathcal{I}_2}\left\{\mathbf{y}^T\mathbf{A}\mathbf{x} + g\|\mathbf{x}\|\|\mathbf{y}\| + \psi(\mathbf{x},\mathbf{y})\right\} \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{I}_1}\max_{\mathbf{y}\in\mathcal{I}_2}\left\{\|\mathbf{x}\|\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y})\right\} \geq c\right)$$

Proof: Define two Gaussian processes indexed on the set $\mathcal{I}_1 \times \mathcal{I}_2$:

$$Y_{\mathbf{x},\mathbf{y}} = \mathbf{x}^T\mathbf{G}\mathbf{y} + g\|\mathbf{y}\|\|\mathbf{x}\| \quad \text{and} \quad X_{\mathbf{x},\mathbf{y}} = \|\mathbf{x}\|\mathbf{g}^T\mathbf{y} - \|\mathbf{y}\|\mathbf{h}^T\mathbf{x}.$$

First, we show that the processes defined satisfy the conditions of Gordon's Theorem A.1. Clearly, they are both centered. Furthermore, for all $\mathbf{x},\mathbf{x}' \in \mathcal{I}_1$ and $\mathbf{y},\mathbf{y}' \in \mathcal{I}_2$:

$$\mathbb{E}[X_{\mathbf{x},\mathbf{y}}^2] = \|\mathbf{x}\|^2\|\mathbf{y}\|^2 + \|\mathbf{y}\|^2\|\mathbf{x}\|^2 = \mathbb{E}[Y_{\mathbf{x},\mathbf{y}}^2],$$

and

$$\begin{aligned}\mathbb{E}[X_{\mathbf{x},\mathbf{y}}X_{\mathbf{x}',\mathbf{y}'}] - \mathbb{E}[Y_{\mathbf{x},\mathbf{y}}Y_{\mathbf{x}',\mathbf{y}'}] &= \|\mathbf{x}\|\|\mathbf{x}'\|(\mathbf{y}^T\mathbf{y}') + \|\mathbf{y}\|^2(\mathbf{x}^T\mathbf{x}') - (\mathbf{x}^T\mathbf{x}')(\mathbf{y}^T\mathbf{y}') - \|\mathbf{y}\|\|\mathbf{y}'\|\|\mathbf{x}\|\|\mathbf{x}'\| \\ &= \left(\underbrace{\|\mathbf{x}\|\|\mathbf{x}'\| - (\mathbf{x}^T\mathbf{x}')}_{\geq 0}\right)\left(\underbrace{(\mathbf{y}^T\mathbf{y}') - \|\mathbf{y}\|\|\mathbf{y}'\|}_{\leq 0}\right),\end{aligned}$$

which is non positive and equal to zero when $\mathbf{x} = \mathbf{x}'$.

Next, for each $(\mathbf{x},\mathbf{y}) \in \mathcal{I}_1 \times \mathcal{I}_2$, let $\lambda_{\mathbf{x},\mathbf{y}} = -\psi(\mathbf{x},\mathbf{y}) + c$ and apply Theorem A.1. This completes the proof by observing that

$$\left[\min_{\mathbf{x}\in\mathcal{I}_1}\max_{\mathbf{y}\in\mathcal{I}_2}\{Y_{\mathbf{x},\mathbf{y}} + \psi(\mathbf{x},\mathbf{y})\} \geq c\right] = \bigcap_{\mathbf{x}\in\mathcal{I}_1}\bigcup_{\mathbf{y}\in\mathcal{I}_2}\left[Y_{\mathbf{x},\mathbf{y}} \geq \lambda_{\mathbf{x},\mathbf{y}}\right],$$

and similar for the process $X_{\mathbf{x},\mathbf{y}}$. ∎

Proof: (of Proposition I.1) Denote $R_1 := \max_{\mathbf{x}\in\mathcal{S}_1}\|\mathbf{x}\|$ and $R_2 := \max_{\mathbf{y}\in\mathcal{S}_2}\|\mathbf{y}\|$. Fix any $\epsilon > 0$. Since $\psi(\cdot,\cdot)$ is continuous and the sets $\mathcal{S}_1$,$\mathcal{S}_2$ are compact, $\psi(\cdot,\cdot)$ is uniformly continuous on $\mathcal{S}_1 \times \mathcal{S}_2$. Thus, there exists $\delta := \delta(\epsilon) > 0$ such that for every $(\mathbf{x},\mathbf{y}),(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in \mathcal{S}_1 \times \mathcal{S}_2$ with $\left\|\begin{bmatrix}\mathbf{x} & \mathbf{y}\end{bmatrix} - \begin{bmatrix}\tilde{\mathbf{x}} & \tilde{\mathbf{y}}\end{bmatrix}\right\| \leq \delta$, we have that $|\psi(\mathbf{x},\mathbf{y}) - \psi(\tilde{\mathbf{x}},\tilde{\mathbf{y}})| \leq \epsilon$. Let $\mathcal{S}_1^\delta$,$\mathcal{S}_2^\delta$ be δ-nets of the sets $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Then, for any $\mathbf{x} \in \mathcal{S}_1$, there exists $\mathbf{x}' \in \mathcal{S}_1^\delta$ such that $\|\mathbf{x} - \mathbf{x}'\| \leq \delta$ and an analogous statement holds for $\mathcal{S}_2$. In what follows, for any vector $\mathbf{v}$ in a set $\mathcal{S}$, we denote $\mathbf{v}'$ the element in the δ-net of $\mathcal{S}$ that is the closest to $\mathbf{v}$ in the usual ℓ_2-metric. To simplify notation, denote

$$\alpha(\mathbf{x},\mathbf{y}) := \mathbf{y}^T\mathbf{A}\mathbf{x} + g\|\mathbf{x}\|\|\mathbf{y}\| + \psi(\mathbf{x},\mathbf{y}) \quad \text{and} \quad \beta(\mathbf{x},\mathbf{y}) := \|\mathbf{x}\|\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y}).$$

From Lemma A.1, we know that for all $c \in \mathbb{R}$:

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y}) \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\beta(\mathbf{x},\mathbf{y}) \geq c\right). \tag{17}$$

In what follows we show that constraining the minimax optimizations over only the δ-nets $\mathcal{S}_1^\delta,\mathcal{S}_2^\delta$ instead of the entire sets $\mathcal{S}_1$,$\mathcal{S}_2$, changes the achieved optimal values by only a small amount.

First, we calculate an upper bound on

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\alpha(\mathbf{x},\mathbf{y}) &\leq \min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y}) =: \alpha(\mathbf{x}_1,\mathbf{y}_1) - \alpha(\mathbf{x}_2,\mathbf{y}_2) \\ &\leq \max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x}_2',\mathbf{y}) - \alpha(\mathbf{x}_2,\mathbf{y}_2) =: \alpha(\mathbf{x}_2',\mathbf{y}_*) - \alpha(\mathbf{x}_2,\mathbf{y}_2) \\ &\leq \alpha(\mathbf{x}_2',\mathbf{y}_*) - \alpha(\mathbf{x}_2,\mathbf{y}_*) \\ &= \mathbf{y}_*^T\mathbf{A}(\mathbf{x}_2' - \mathbf{x}_2) + g\|\mathbf{y}_*\|(\|\mathbf{x}_2'\| - \|\mathbf{x}_2\|) + (\psi(\mathbf{x}_2',\mathbf{y}_*) - \psi(\mathbf{x}_2,\mathbf{y}_*)) \\ &\leq (\|\mathbf{A}\|_2 + |g|)\underbrace{\|\mathbf{y}_*\|}_{\leq R_2}\underbrace{\|\mathbf{x}_2' - \mathbf{x}_2\|}_{\leq\delta} + \underbrace{|\psi(\mathbf{x}_2',\mathbf{y}_*) - \psi(\mathbf{x}_2,\mathbf{y}_*)|}_{\leq\epsilon} \\ &\leq (\|\mathbf{A}\|_2 + |g|)R_2\delta + \epsilon.\end{aligned}$$

From this, we have that

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\alpha(\mathbf{x},\mathbf{y})\geq c\right)\geq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y})\geq c+(\|\mathbf{A}\|_2+|g|)R_2\delta+\epsilon\right). \tag{18}$$

Using standard concentration results on Gaussians, it is shown in Lemma B.1 that for all $t>0$,

$$\mathbb{P}(\|\mathbf{A}\|_2+|g|\leq\sqrt{m}+\sqrt{n}+1+t)\geq 1-2\exp(-t^2/4).$$

This, when combined with (18) yileds:

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\alpha(\mathbf{x},\mathbf{y})\geq c\right)\geq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\alpha(\mathbf{x},\mathbf{y})\geq c+(\sqrt{n}+\sqrt{m}+1+t)R_2\delta+\epsilon\right)-2\exp(-t^2/4). \tag{19}$$

Similarly,

$$\begin{aligned}
\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\beta(\mathbf{x},\mathbf{y})-\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})&\geq\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\beta(\mathbf{x},\mathbf{y})-\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})=:\beta(\mathbf{x}_1,\mathbf{y}_1)-\beta(\mathbf{x}_2,\mathbf{y}_2)\\
&\geq\beta(\mathbf{x}_1,\mathbf{y}_1)-\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x}_1,\mathbf{y})=:\beta(\mathbf{x}_1,\mathbf{y}_1)-\beta(\mathbf{x}_1,\mathbf{y}_*)\\
&\geq\beta(\mathbf{x}_1,\mathbf{y}_*')-\beta(\mathbf{x}_1,\mathbf{y}_*)\\
&=\|\mathbf{x}_1\|\mathbf{g}^T(\mathbf{y}_*'-\mathbf{y}_*)+(\|\mathbf{y}_*'\|-\|\mathbf{y}_*\|)\mathbf{h}^T\mathbf{x}_1+(\psi(\mathbf{x}_1,\mathbf{y}_*')-\psi(\mathbf{x}_1,\mathbf{y}_*))\\
&\geq-(\|\mathbf{g}\|+\|\mathbf{h}\|)\underbrace{\|\mathbf{x}_1\|}_{\leq R_1}\underbrace{\|\mathbf{y}_*'-\mathbf{y}_*\|}_{\leq\delta}-\underbrace{|\psi(\mathbf{x}_1,\mathbf{y}_*')-\psi(\mathbf{x}_1,\mathbf{y}_*)|}_{\leq\epsilon}\\
&\geq-(\|\mathbf{g}\|+\|\mathbf{h}\|)R_1\delta-\epsilon.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})\geq c+(\|\mathbf{g}\|+\|\mathbf{h}\|)R_1\delta+\epsilon\right)\leq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\beta(\mathbf{x},\mathbf{y})\geq c\right),$$

and a further application of Lemma B.1 shows that for all $t>0$:

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})\geq c+(\sqrt{n}+\sqrt{m}+t)R_2\delta+\epsilon\right)-2\exp(-t^2/4)\leq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1^\delta}\max_{\mathbf{y}\in\mathcal{S}_2^\delta}\beta(\mathbf{x},\mathbf{y})\geq c\right), \tag{20}$$

Now, we can apply (17) in order to combine (19) and (20) to yield the following:

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\alpha(\mathbf{x},\mathbf{y})\geq c\right)\geq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})\geq c+(\sqrt{n}+\sqrt{m}+1+t)(R_1+R_2)\delta+2\epsilon\right)-4\exp(-t^2/4).$$

This holds for all $\epsilon>0$ and all $t>0$. In particular, set $t=\delta^{-\frac{1}{2}}$ and take the limit of the right-hand side as $\epsilon\to 0$. Then, $t\to\infty$ and we can of course choose $\delta\to 0$, which proves that

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\alpha(\mathbf{x},\mathbf{y})\geq c\right)\geq\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\beta(\mathbf{x},\mathbf{y})>c\right).$$

■

B. Auxiliary Results

Definition B.1 (Lipschitz)**.** *We say that a function $f:\mathbb{R}^d\to\mathbb{R}$ is Lipschitz with constant L or is L-Lipschitz if $|f(\mathbf{x})-f(\mathbf{y})|\leq L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.*

Proposition B.1 (Gaussian Lipschitz concentration)**.** *([22, Theorem 5.6]) Let $\mathbf{x}\in\mathbb{R}^d$ have i.i.d. $\mathcal{N}(0,1)$ entries and $f:\mathbb{R}^d\to\mathbb{R}$ be L-Lipschitz. Then, each one of the events $\{f(\mathbf{x})>\mathbb{E}f(\mathbf{x})+t\}$ and $\{f(\mathbf{x})<\mathbb{E}f(\mathbf{x})-t\}$ occurs with probability no greater than $\exp\left(-t^2/(2L^2)\right)$.*

Lemma B.1. *Let $\mathbf{A}\in\mathbb{R}^{m\times n}$, $g\in\mathbb{R}$, $\mathbf{g}\in\mathbb{R}^m$ and $\mathbf{h}\in\mathbb{R}^n$ have entries i.i.d. $\mathcal{N}(0,1)$ and be independent of each other. Then, for all $t>0$, each one of the events*

$$\{\|\mathbf{A}\|_2+|g|\leq\sqrt{n}+\sqrt{m}+1+t\}\quad\text{and}\quad\{\|\mathbf{h}\|_2+\|\mathbf{g}\|_2\leq\sqrt{n}+\sqrt{m}+t\}, \tag{21}$$

holds with probability at least $1-2\exp(-t^2/4)$.

Proof: A well-known non-asymptotic bound on the largest singular value of an $m\times n$ Gaussian matrix shows (e.g. [23, Corollary 5.35]) that for all $t>0$:

$$\mathbb{P}\left(\|\mathbf{A}\|_2>\sqrt{m}+\sqrt{n}+t\right)\leq\exp(-t^2/2).$$

Also, $\|\cdot\|_2$ is an 1-Lipschitz function and for a standard gaussian vector $\mathbf{v}\in\mathbb{R}^d$: $\mathbb{E}\|\mathbf{v}\|_2\leq\sqrt{d}$. Applying Proposition B.1 we have that for all $t>0$ the events $\{|g|>1+t\}$, $\{\|\mathbf{g}\|_2>\sqrt{m}+t\}$ and $\{\|\mathbf{h}\|_2>\sqrt{n}+t\}$, each one occurs with probability no larger than $\exp(-t^2/2)$. Combining those,

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{A}\|_2+|g|\leq\sqrt{n}+\sqrt{m}+1+t\right)&\geq\mathbb{P}\left(\|\mathbf{A}\|_2\leq\sqrt{n}+\sqrt{m}+t/2\ ,\ |g|\leq 1+t/2\right)\\&\geq 1-\mathbb{P}\left(\|\mathbf{A}\|_2>\sqrt{n}+\sqrt{m}+t/2\right)-\mathbb{P}\left(\ |g|>1+t/2\right)\\&\geq 1-2\exp(-t^2/4).\end{aligned}$$

The proof of the second statement of the lemma is identical and is omitted for brevity. ■

Lemma B.2 (Lipschitzness of Gordon's Optimization)**.** *Let $\mathcal{S}_1\subset\mathbb{R}^n$, $\mathcal{S}_2\subset\mathbb{R}^m$ be compact sets and function $\mathcal{G}:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$:*

$$\mathcal{G}(\mathbf{g},\mathbf{h}):=\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\ \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y}+\|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x}+\psi(\mathbf{x},\mathbf{y}).$$

Further let $R_1=\max_{\mathbf{x}\in\mathcal{S}_1}\|\mathbf{x}\|_2$ and $R_2=\max_{\mathbf{y}\in\mathcal{S}_2}\|\mathbf{y}\|_2$. Then, $\mathcal{G}(\mathbf{g},\mathbf{h})$ is Lipschitz with constant $\sqrt{2}R_1R_2$.

Proof: Fix any two pairs $(\mathbf{g}_1,\mathbf{h}_1)$ and $(\mathbf{g}_2,\mathbf{h}_2)$ and let

$$(\mathbf{x}_2,\mathbf{y}_2)=\arg\min_{\mathbf{x}\in\mathcal{S}_1}\max_{\mathbf{y}\in\mathcal{S}_2}\|\mathbf{x}\|\mathbf{g}_2^T\mathbf{y}+\|\mathbf{y}\|\mathbf{h}_2^T\mathbf{x}+\psi(\mathbf{x},\mathbf{y}),$$

and

$$\mathbf{y}_*=\arg\max_{\mathbf{y}\in\mathcal{S}_2}\|\mathbf{x}_2\|\mathbf{g}_1^T\mathbf{y}+\|\mathbf{y}\|\mathbf{h}_1^T\mathbf{x}_2+\psi(\mathbf{x}_2,\mathbf{y}).$$

Clearly,

$$\mathcal{G}(\mathbf{g}_1,\mathbf{h}_1)\leq\|\mathbf{x}_2\|\mathbf{g}_1^T\mathbf{y}_*+\|\mathbf{y}_*\|\mathbf{h}_1^T\mathbf{x}_2+\psi(\mathbf{x}_2,\mathbf{y}_*),$$

and

$$\mathcal{G}(\mathbf{g}_2,\mathbf{h}_2)\geq\|\mathbf{x}_2\|\mathbf{g}_2^T\mathbf{y}_*+\|\mathbf{y}_*\|\mathbf{h}_2^T\mathbf{x}_2+\psi(\mathbf{x}_2,\mathbf{y}_*),$$

Without loss of generality, assume $\mathcal{G}(\mathbf{g}_1,\mathbf{h}_1)\geq\mathcal{G}(\mathbf{g}_2,\mathbf{h}_2)$. Then,

$$\begin{aligned}\mathcal{G}(\mathbf{g}_1,\mathbf{h}_1)-\mathcal{G}(\mathbf{g}_2,\mathbf{h}_2)&\leq\|\mathbf{x}_2\|\mathbf{g}_1^T\mathbf{y}_*+\|\mathbf{y}_*\|\mathbf{h}_1^T\mathbf{x}_2+\psi(\mathbf{x}_2,\mathbf{y}_*)-(\|\mathbf{x}_2\|\mathbf{g}_2^T\mathbf{y}_*+\|\mathbf{y}_*\|\mathbf{h}_2^T\mathbf{x}_2+\psi(\mathbf{x}_2,\mathbf{y}_*))\\&\leq\|\mathbf{x}_2\|\mathbf{y}_*^T(\mathbf{g}_1-\mathbf{g}_2)+\|\mathbf{y}_*\|\mathbf{x}_2^T(\mathbf{h}_1-\mathbf{h}_2)\\&\leq\sqrt{\|\mathbf{x}_2\|^2\|\mathbf{y}_*\|^2+\|\mathbf{y}_*\|^2\|\mathbf{x}_2\|^2}\sqrt{\|\mathbf{g}_1-\mathbf{g}_2\|^2+\|\mathbf{h}_1-\mathbf{h}_2\|^2}\\&\leq R_1R_2\sqrt{2}\sqrt{\|\mathbf{g}_1-\mathbf{g}_2\|^2+\|\mathbf{h}_1-\mathbf{h}_2\|^2},\end{aligned}$$

where the penultimate inequality follows from Cauchy-Schwarz. ■

C. Gordon's Optimization for the LASSO Objective

In this section we formalize the discussion of Section IV and prove (15). The main result is stated in Lemma C.1. Some auxiliary results required for the proof of C.1 are presented as separate lemmas.

Lemma C.1. *Let $\mathcal{G}(\mathbf{g},\mathbf{h},h):\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$ be defined as in (14) and $\mathbf{g}$,$\mathbf{h}$ and h have entries i.i.d. standard normal. Assume $(1-\epsilon)\gamma_m>\omega>\epsilon\gamma_m$ for some $\epsilon>0$. Further assume*

$$C>C_*:=\sigma\frac{\omega}{\sqrt{\gamma_m^2-\omega^2}}.$$

Then,
(a) $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)\leq\sigma\sqrt{\gamma_m^2-\omega^2}+\sqrt{C^2+\sigma^2}\sqrt{\pi}$.
(b) For all $0<\delta<1$, there exists sufficiently large m, such that $\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)\geq(1-\delta)\sigma\sqrt{\gamma_m^2-\omega^2}$.

Proof:
(a) Denote $\mathcal{G}_1(\mathbf{g},\mathbf{h},h)=\min_{0\leq\alpha\leq C}\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2-\alpha d(\mathbf{h})\}+h\sigma$. Then,

$$\begin{aligned}\mathbb{E}\mathcal{G}_1(\mathbf{g},\mathbf{h},h)\leq\min_{0\leq\alpha\leq C}\mathbb{E}\left[\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2-\alpha d(\mathbf{h})\right]&=\min_{0\leq\alpha\leq C}\sqrt{\alpha^2+\sigma^2}\gamma_m-\alpha\omega\\&=\sigma\sqrt{\gamma_m^2-\omega^2}.\end{aligned}\tag{22}$$

The last equality above follows from Lemma C.2 and the assumptions $\gamma_m>\omega$ and $C>C_*$. In Lemma C.3 we show that $\mathcal{G}_1(\mathbf{g},\mathbf{h},h)$ is Lipschitz with constant $\sqrt{2}\sqrt{C^2+\sigma^2}$. Thus, we can apply Lemma C.4 and combine with (22) to prove the desired:

$$\begin{aligned}\mathbb{E}\mathcal{G}(\mathbf{g},\mathbf{h},h)=\mathbb{E}\left(\mathcal{G}_1(\mathbf{g},\mathbf{h},h)\right)_+&\leq\left(\mathbb{E}\mathcal{G}_1(\mathbf{g},\mathbf{h},h)\right)_+\sqrt{C^2+\sigma^2}\sqrt{\pi}\\&\leq\sigma\sqrt{\gamma_m^2-\omega^2}+\sqrt{C^2+\sigma^2}\sqrt{\pi}.\end{aligned}$$

(b) Fix any $0 < \delta < 1$. Let $\eta := \min\{\delta/2\ ,\ \epsilon/(1-\epsilon)\}$ and define the event

$$\mathcal{E} := \{\|\mathbf{g}\| \geq \gamma_m - \eta(\gamma_m - \omega) \quad \text{and} \quad |d(\mathbf{h}) - \omega| \leq \eta(\gamma_m - \omega)\}\,.$$

Clearly, from the non-negativity of $\mathcal{G}(\mathbf{g}, \mathbf{h}, h)$:

$$\mathbb{E}\mathcal{G}(\mathbf{g}, \mathbf{h}, h) \geq \mathbb{E}\left[\mathcal{G}(\mathbf{g}, \mathbf{h}, h)\Big|\mathcal{E}\right]\mathbb{P}(\mathcal{E}). \tag{23}$$

In what follows, we calculate $\mathbb{P}(\mathcal{E})$ and $\mathbb{E}[\mathcal{G}(\mathbf{g}, \mathbf{h}, h)|\mathcal{E}]$ and use (23) to complete the proof.

First, we show that $\mathcal{E}$ holds with high probability. Both functions $\|\cdot\|_2$ and $d(\cdot)$ are 1-Lipschitz. Applying Proposition B.1:

$$\mathbb{P}\left(|d(\mathbf{h}) - \omega| > \eta(\gamma_m - \omega)\right) \leq 2\exp(-\eta^2(\gamma_m - \omega)^2/2),$$

and

$$\mathbb{P}(\|\mathbf{g}\| < \gamma_m - \eta(\gamma_m - \omega)) \leq \exp(-\eta^2(\gamma_m - \omega)^2/2).$$

Combining these with a union bound and using $\gamma_m - \omega > \epsilon\gamma_m$:

$$\mathbb{P}\left(\mathcal{E}\right) \geq 1 - 3e^{-\eta^2\epsilon^2\gamma_m^2/2}. \tag{24}$$

Next, we compte $\mathbb{E}\left[\mathcal{G}(\mathbf{g}, \mathbf{h}, h)\Big|\mathcal{E}\right]$. Observe that

$$\mathcal{G}(\mathbf{g}, \mathbf{h}, h) \geq \min_{\alpha \geq 0}\{\sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \alpha d(\mathbf{h})\} + h\sigma.$$

Hence,

$$\mathbb{E}\left[\mathcal{G}(\mathbf{g}, \mathbf{h}, h)\Big|\mathcal{E}\right] \geq \mathbb{E}\left[\min_{\alpha \geq 0}\{\sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \alpha d(\mathbf{h})\} + h\sigma\Big|\mathcal{E}\right] = \mathbb{E}\left[\min_{\alpha \geq 0}\{\sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \alpha d(\mathbf{h})\}\Big|\mathcal{E}\right], \tag{25}$$

where the last equation follows since h is independent of $\mathbf{g}, \mathbf{h}$ and zero-mean. It only takes a few algebra steps to show that when $\mathcal{E}$ holds, the following are true:

$$\|\mathbf{g}\|_2^2 - d(\mathbf{h})^2 \geq (1 - 2\eta)(\gamma_m^2 - \omega^2) > 0,$$
$$d(\mathbf{h}) \geq \omega - \eta(\gamma_m - \omega) > ((1 + \eta)\epsilon - \eta)\gamma_m \geq 0.$$

In the above, we have also used the facts that $\eta < 1/2$ and $\omega > \epsilon\gamma_m$. Thus, we may apply Lemma C.2 to find that

$$(25) = \sigma\mathbb{E}\left[\sqrt{\|\mathbf{g}\|^2 - d(\mathbf{h})^2}\Big|\mathcal{E}\right] \geq \sigma\sqrt{1 - 2\eta}\sqrt{\gamma_m^2 - \omega^2}.$$

This when combined with (24) and (23) yields:

$$\mathbb{E}\mathcal{G}(\mathbf{g}, \mathbf{h}, h) \geq \sqrt{1 - 2\eta}\left(1 - 3e^{-\eta^2\epsilon^2\gamma_m^2/2}\right)\sigma\sqrt{\gamma_m^2 - \omega^2}. \tag{26}$$

Observe that $\sqrt{1 - 2\eta} \geq \sqrt{1 - \delta} > 1 - \delta$. Thus, we can choose m large enough in (26) to complete the proof. ■

Lemma C.2 (Scalar Optimization)**.** *Let $b > d > 0$, $\sigma > 0$ and $C > \sigma\frac{d}{\sqrt{b^2 - d^2}}$. Then,*

$$\min_{0 \leq \alpha \leq C}\left\{\sqrt{\alpha^2 + \sigma^2}\ b - \alpha d\right\} = \sigma\sqrt{b^2 - d^2}.$$

Proof: Consider the relaxed minimization problem $\min_{0 \leq \alpha} f(\alpha)$ where $f(\alpha) = \sqrt{\alpha^2 + 1}\ b - \alpha d$. We easily calculate the derivative of the objective $f'(\alpha) = \frac{\alpha}{\sqrt{\alpha^2 + \sigma^2}}b - d$. If $b > d > 0$, then $\alpha_* = \sigma d/\sqrt{b^2 - d^2} > 0$ sets $f'(\cdot)$ equal to zero; thus, is optimal and a straightforward calculation yields $f(\alpha_*) = \sigma\sqrt{b^2 - d^2}$. ■

Lemma C.3. *The function $\mathcal{G}_1 : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ defined as*

$$\mathcal{G}_1(\mathbf{g}, \mathbf{h}, h) = \min_{0 \leq \alpha \leq C}\{\sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \alpha d(\mathbf{h})\} + h\sigma,$$

is Lipschitz with Lipschitz constant $\sqrt{2}\sqrt{C^2 + \sigma^2}$.

Proof: Consider $(\mathbf{g}_1, \mathbf{h}_1, h_1)$ and $(\mathbf{g}_2, \mathbf{h}_2, h_2)$. Without loss of generality, assume $\mathcal{G}_1(\mathbf{g}_1, \mathbf{h}_1, h_1) > \mathcal{G}_1(\mathbf{g}_2, \mathbf{h}_2, h_2)$ and let α_2 be the optimal value achieving $\mathcal{G}_1(\mathbf{g}_2, \mathbf{h}_2, h_2)$. Then,

$$\begin{aligned}
\mathcal{G}(\mathbf{g}_1, \mathbf{h}_1, h_1) - \mathcal{G}(\mathbf{g}_2, \mathbf{h}_2, h_2) &\leq \sqrt{\alpha_2^2 + \sigma^2}\|\mathbf{g}_1\|_2 - \alpha_2 d(\mathbf{h}_1) + h_1\sigma - \left(\sqrt{\alpha_2^2 + \sigma^2}\|\mathbf{g}_2\|_2 - \alpha_2 d(\mathbf{h}_2) + h_2\sigma\right) \\
&\leq \sqrt{\alpha_2^2 + \sigma^2}(\|\mathbf{g}_1\|_2 - \|\mathbf{g}_2\|_2) + \alpha_2(d(\mathbf{h}_2) - d(\mathbf{h}_1)) + \sigma(h_1 - h_2) \\
&\leq \sqrt{\alpha_2^2 + \sigma^2}(\|\mathbf{g}_1 - \mathbf{g}_2\|_2) + \alpha_2\|\mathbf{h}_2 - \mathbf{h}_1\|_2 + \sigma|h_1 - h_2| \\
&\leq \sqrt{2(\alpha_2^2 + \sigma^2)}\sqrt{\|\mathbf{g}_1 - \mathbf{g}_2\|_2^2 + \|\mathbf{h}_1 - \mathbf{h}_2\|_2^2 + (h_1 - h_2)^2}, \qquad (27)
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. Use the constraint $\alpha_2 \leq C$ to conclude the proof. ■

Lemma C.4. *Let $\mathbf{x} \in \mathbb{R}^n$ have standard normal entries and $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. Then,*

$$\mathbb{E}f(\mathbf{x}) \leq \mathbb{E}\left(f(\mathbf{x})\right)_+ \leq \left(\mathbb{E}f(\mathbf{x})\right)_+ + L\sqrt{\pi/2}$$

Proof: The left-hand side of the claimed inequality is straightforward. In what follows, we prove the right-hand side. There are two cases to consider.

Case 1: Suppose $\mathbb{E}f(\mathbf{x}) \geq 0$. Observe that

$$f(\mathbf{x}) = \left(f(\mathbf{x})\right)_+ + \min\{0, f(\mathbf{x})\}. \qquad (28)$$

We will bound $\mathbb{E}\min\{0, f(\mathbf{x})\}$:

$$\begin{aligned}
\mathbb{E}\min\{0, f(\mathbf{x})\} = -\int_{-\infty}^{0} \mathbb{P}(f(\mathbf{x}) \leq c)\mathrm{d}c &= -\int_{\mathbb{E}f(\mathbf{x})}^{\infty} \mathbb{P}(f(\mathbf{x}) \leq \mathbb{E}f(\mathbf{x}) - t)\mathrm{d}t \\
&\geq -\int_{\mathbb{E}f(\mathbf{x})}^{\infty} \exp(-t^2/(2L^2))\mathrm{d}t \qquad (29) \\
&= -L\int_{\mathbb{E}f(\mathbf{x})/L}^{\infty} \exp(-u^2/2)\mathrm{d}u \\
&\geq -L\int_{0}^{\infty} \exp(-u^2/2)\mathrm{d}u \geq -L\sqrt{\pi/2}. \qquad (30)
\end{aligned}$$

(29) follows from the Lipschitzness assumption and Proposition B.1. Combining (30) with (28) proves the desired.

Case 2: Suppose $\mathbb{E}f(\mathbf{x}) < 0$. Observe that $\min\{0, -f(\mathbf{x})\} = -\left(f(\mathbf{x})\right)_+$. Apply (30) to conclude with $\mathbb{E}\left(f(\mathbf{x})\right)_+ \leq L\sqrt{\pi/2}$. ■